MARTIN DILLON
Director
ERIK JUL
Communications Manager
MARK BURGE
Research Assistant
CAROL HICKEY
Research Assistant

OCLC Online Computer Library Center, Inc.
Office of Research
Dublin, Ohio

The OCLC Internet Resources Project: Toward Providing Library Services for Computer-Mediated Communication

INTRODUCTION

Locating, accessing, and using information resources on the Internet, a global computer network of networks, can be difficult, time-consuming, and sometimes impossible. In this new and rapidly expanding electronic environment, network users have unprecedented access to information and computing resources. However, the development and implementation of systematic methods of describing and accessing information lag behind deployment of the Internet itself. Network users' ability to share information surpasses by far their ability to discover information on the Internet. Traditional library services such as cataloging have yet to find widespread application in this emerging environment.

Funded by the U.S. Department of Education, Library Programs, the OCLC Internet Resources project investigated the nature of electronic textual information accessible via the Internet. This empirical study also explored the practical and theoretical problems associated with providing traditional library services for electronic text in a wide area network environment. This report presents the findings and recommendations arising from the project, as well as suggested areas for further study.

OBJECTIVES

The primary objectives of this project were, first, to provide an empirical analysis of textual information on the Internet; second, to test the suitability of current cataloging rules and record formats governing the creation of machine-readable cataloging records; and, third, to develop recommendations that would assist the efforts of standards bodies and others interested in systematically cataloging or otherwise describing and providing access to electronic information objects available through remote network access.

METHODS

Project methods included, first, locating, collecting, and analyzing a sample of textual information objects derived from sources accessible via the Internet; second, developing and testing a taxonomy of electronic information based on the sample; and, third, conducting a cataloging experiment to identify and analyze problems associated with cataloging and providing appropriate levels of access to this information.

REFERENCE SOURCES

The early focus of the project was to collect sample text documents from Internet sources. In fall 1991, when the project was initiated, few resources existed to describe or assist access to Internet resources. Primary source materials initially available to project staff included Barron (1992), Frey and Adams (1991), Malkin and Marine (1992), National Science Foundation (1989), Krol (1989), LaQuey (1990), Quarterman (1990), and St. George and Larsen (1991). A now-defunct, manually maintained list of Internet File Transfer Protocol (FTP) sites also provided initial direction. The "Request for Comments" series of documents, which form the official Internet documentation sanctioned by the Internet Activities Board, also provided valuable initial direction. In the year of the study, print and electronic guides, directories, and other reference materials proliferated, and general discussion of the Internet moved from government, technical, or trade publications to the popular press.

Not surprisingly, much of the information about Internet resources is published, at least originally, in electronic form for distribution across the network. For the novice network user and those without Internet access, this is a hindrance to knowledge and a source of frustration. In recognition of this problem, several quality users' guides have been published recently in traditional book form (see, e.g., Kehoe, 1993; Krol, 1992; and Marine, 1992).

In addition to print and electronic reference tools, project staff used an array of systems specifically designed to assist the discovery and access of Internet resources. These electronic aids included WAIS (Wide Area Information Servers) by Thinking Machines Corporation, Gopher by the University of Minnesota, archie by McGill University, HYTELNET by Peter Scott, and electronic conferences (Kovacs, 1991; Strangelove, Okerson, & Kovacs, 1992).

WAIS is a distributed search and retrieval system using a client/server model and the draft Z39.50 standard for bibliographic retrieval (see Kahle & Medlar, 1991; Lincoln, 1992a, 1992b; and Nickerson, 1992). Gopher is a client/server protocol for distributed information systems (see Alberti, Anklesaria, Lindner, McCahill, & Torrey, 1992). The archie system facilitates information discovery and access by creating a searchable database of file and directory information obtained from FTP host computers (see Deutsch, 1992). HYTELNET provides hypertext access to lists of Internet resources and facilitates logging on to the remote resource (see Scott, 1992). These methods were augmented by electronic mail and online browsing.

ANALYSIS OF FTP SITES

The TCP/IP (Transmission Control Protocol/Internet Protocol) protocol suite provides FTP, allowing the transfer of electronic files among remote computers. Using FTP, system administrators can designate computers as anonymous FTP servers, that is, computers allowing anonymous FTP access to a store of files.

A feature of this protocol allows users to log on to remote computers on which they do not have an account as an "anonymous" user. Users can traverse the computer's file structure, display directory and file names, and initiate the transfer of files from or to the remote site. FTP prevents users from accessing other portions of the computer's file system.

FTP accounts for a significant portion of network traffic (46% of characters carried by the network as compared to 18% for electronic mail, 6% for Telnet, and 29% for other traffic). Internet traffic statistics derive from various sources and represent a snapshot of network development at a particular point in time (see Lottor, 1992, and network statistics provided by the Merit Network, Inc., available by FTP NIC.MERIT.EDU; directory: nsfnet/statistics; file: history.netcount). For this reason, project staff undertook a detailed analysis of FTP sites.

Method

Investigation of electronic documents was undertaken through manual collection and analysis and through computer-assisted statistical analyses and automated categorization. Each of these methods is described in the following sections.

File Collection and Analysis

The early focus of the project was to collect sample text documents from Internet sources. Project staff used an array of resources to discover the whereabouts of electronic text, including printed books, journal articles, and newsletters; online electronic publications and lists; information discovery tools such as WAIS, Gopher, and archie; hypertext programs; electronic conferences; electronic mail; and online browsing.

Project staff sought to categorize and quantify the information available via FTP sites automatically. This investigation was facilitated by data collected

by the archie service, developed by Peter Deutsch and Alan Emtage of McGill University. The archie service is an early entry into the field of wide area information discovery. In short, the archie service has developed software that attempts to discover anonymous FTP sites and their contents. The software initiates an anonymous FTP logon at Internet host sites, cycling through the entire list of sites approximately once every 30 days. If the anonymous FTP logon is successful, the software executes a listing of the FTP site's directories, thus obtaining a list of every available file at the site. The file names extracted from the FTP sites are stored in a file and mounted in a searchable database. Users of the archie service can search the database for file names, and the system will provide the Internet address for sites containing files whose names match the user's query.

The archie service is a ready source of information about FTP sites and provides data that served as a starting point for generating a statistical snapshot of Internet resources. The file containing the list of FTP sites and their holdings is itself available via FTP, and project staff obtained it to extract a database for processing and analysis. The database includes a listing of FTP sites, paths, names, and file sizes, along with several other pieces of informaiton about each site.

To discover trends in the growth of FTP sites, we created and analyzed this database periodically. This sampling revealed rapid growth in the number of FTP sites during the time of this study, the number of files available at these sites, and the amount of data stored on magnetic disk. From January 1992 to August 1992, the number of sites grew from 829 to 1,044, a 25.93% increase; the number of files grew from 2,089,544 to 3,059,689, a 46.43% increase; and the size increased from 101.02 Gbytes to 165.05 Gbytes, a 63.38% increase.

To begin to get a sense of the makeup of these FTP sites, we selected 20 sites at random for closer analysis (Table 1). This sample clearly shows a wide range of profiles by every measure, including the number of files at a site (from 12 to 38,440), the amount of data stored (from 104,969 to 913,679,044 bytes), the largest file (from 45,056 to 28,437,472 bytes), and the average file size (8,747 to 2,530,930 bytes). The distribution of data among the sample sites is uneven; for example, the site csam.lbl.gov has only 57 files, yet it contains the largest file in the sample (28,437,472 bytes) and has the largest average file size (2,530,930 bytes). In contrast, the largest site in the sample, lth.se, has both the most files (38,440) and the most storage (913,679,044 bytes) but a comparatively low average file size (23,768).

The 20 largest FTP sites are shown in Table 2. At the time of our sampling, the largest FTP site on the Internet in terms of total files and total storage was src.doc.ic.ac.uk; the largest file, "db.pag" (1,846,821,888 bytes), was also at this site. These 20 largest sites, or 2% of Internet FTP sites, account for 57% of the available files and 38% of the storage, again revealing the disproportionate distribution of data and the significant contribution made by several large sites.

Directories, Paths, and File Names

Collectively, the Internet's anonymous FTP sites may be viewed as an archive or "library" of electronic information. Project staff investigated the methods

TABLE 1
SURVEY OF 20 SAMPLE FTP SITES BY NUMBER OF FILES

Site Name	*No. Files*	*Total Bytes*	*Largest File (Bytes)*	*Average File (Bytes)*
lth.se	38,440	913,679,044	13,344,768	23,768
research.att.com	9,102	257,800,968	8,752,643	28,323
archive.cis.ohio-state.edu	8.843	669,287,526	7,287,625	75,685
merit.edu	1,696	147,797,681	2,546,131	87,144
ftp.cica.indiana.edu	1,475	167,161,346	2,052,422	113,329
hubcap.clemson.edu	726	75,764,452	5,455,054	104,358
a.cs.uiuc.edu	459	65,460,295	6,097,773	142,615
turbo.bio.net	390	16,851,476	750,368	43,208
boombox.micro.umn.edu	382	36,239,511	2,047,933	94,867
dsl.cis.upenn.edu	134	6,102,098	816,261	45,538
gem.stack.urc.tue.nl	142	13,309,519	1,081,976	93,729
okeeffe.cs.berkeley.edu	115	18,078,337	3,853,003	157,202
nic.mr.net	124	9,552,240	1,854,848	77,034
watcgl.waterloo.edu	124	3,352,867	1,033,077	27,039
shemp.cs.ucla.edu	58	8,393,452	1,696,416	144,714
csam.lbl.gov	57	144,263,025	28,437,472	2,530,930
paul.rutgers.edu	19	1,638,718	602,699	86,248
suna.osc.edu	18	1,574,391	342,822	87,466
uop.uop.edu	16	4,174,683	784,987	260,917
jhname.hcf.jhu.edu	12	104,969	45,056	8,747

currently used to classify, describe, and facilitate the location of and access to information at these sites.

Apart from the FTP site names, other indicators of the type and location of information available at the site include the directory names, path names (a hierarchical series of directory names), and the individual file names. Minimally, any particular file will have a file name, directory name, and site name associated with it. Each of these names may provide meaningful information about the nature and contents of a file. In aggregate, these names may produce a cogent hierarchy of descriptors or they may be unintelligible to anyone but the creator of the directory/path/file-name structure.

The directory and path names provide description and location information for the files contained at the FTP site. On average, a typical file has fewer than three (2.47) associated content/location indicators, including the file name itself (Table 3). This indicates that the average hierarchical file structure is rather shallow and may provide inadequate descriptive information. (The depth of a hierarchical file structure does not affect the utility of location information.)

Readme and Index Files

FTP sites may contain text files that provide additional descriptive information about the contents of the site, a particular directory, or particular files. These informational files are often named "readme," "index," or some variation thereof in combination with other characters. The value of these informational files can vary greatly depending on the completeness, clarity, and currency of the descriptive information provided.

TABLE 2
20 LARGEST FTP SITES BY NUMBER OF FILES*

Site Name	*No. Files*	*Total Bytes*	*Largest File*	*Average File*
src.doc.ic.ac.uk	170,966	7,923,289,150	1,846,821,888**	46,344
wuarchive.wustl.edu	147,173	6,039,051,548	30,121,209	41,033
capella.eetech.mcgill.ca	131,262	5,199,556,552	30,121,209	39,612
mcsun.eu.net	109,483	1,065,088,972	12,082,830	9,728
isfs.kuis.kyoto-u.ac.jp	76,880	4,022,047,707	24,169,327	52,315
ucs.edu	67,288	289,291,834	12,886,016	4,299
gatekeeper.dec.com	67,100	4,279,830,040	44,877,484	63,782
toklab.ics.osaka-u.ac.jp	65,135	2,237,389,271	25,518,080	34,350
ftp.uu.net	59,508	2,689,716,008	10,573,106	45,199
plaza.aarnet.edu.au	54,046	3,677,983,744	30,121,209	68,052
athene.uni-paderborn.de	49,418	2,486,320,000	11,534,336	50,312
stis.nsf.gov	40,792	102,695,505	5,124,940	2,517
emx.cc.utexas.edu	40,550	478,590,134	4,841,472	11,802
erratic.bradley.edu	38,687	987,391,765	5,458,229	25,522
ipcl.rrzn.uni-hannover.de	38,511	1,291,990,465	33,144,095	33,548
lth.se	38,440	913,679,044	13,344,768	23,768
faui43.informatik.uni-erlangen.de	35,091	2,214,839,896	12,881,920	63,117
cs.ubc.ca	33,744	1,460,556,438	20,200,637	43,283
arp.anu.edu.au	32,142	126,915,618	2,803,093	3,948
rusmv1.rus.uni-stuttgart.de	28,963	1,573,641,267	46,097,964	54,332

* The top 20 Internet sites account for 57% of the available files and 38% of the storage.
** Largest file at Internet FTP site: /ic.doc/whois/db.pag.

TABLE 3
DIRECTORIES AT FTP SITES

Sites	1,044.00
Files	3,059,689.00
Directories	192,446.00
File/directory (avg.)	15.90
Directories/site (avg.)	184.34
Top-level directories	4,861.00
Top-level directories/site (avg.)	4.66
Maximum directory nodes*	20.00
Average directory nodes/file*	2.47

* Number of nodes in directory path including file name (std 1.26).

Project staff examined the frequency of these auxiliary informational files in a 20-site sample. Based on this sample, there is one readme file for every 3.5 directories and one index file for every 7 directories. By extrapolation, there is one readme file for every 55.65 files and one index file for every 111.3 files. Thus, despite the potential utility of these files, they occur infrequently.

TYPES OF FILES AT FTP SITES

Acquiring a statistical overview of FTP sites is useful and straightforward; determining the contents of FTP sites is more difficult and, for the average user, more necessary. Project staff sought to determine the composition of the aggregate FTP sites using automated methods. The chief and most readily available guides to the nature and contents of files at FTP sites are the directory and file names. Drawing from the 20-site sample, project staff compiled a list of all path names (the complete hierarchical path for each file in the data set), which were then counted and sorted by frequency. (A directory name was counted each time it occurred in a path for a file. For example, many FTP sites organize publicly accessible files hierarchically under the top-level directory "pub." Thus, while the /pub directory may have the most associated files, it likely occurs only once in file hierarchy at any given site.) The directory names in this sample set are highly idiosyncratic but nevertheless representative of the type of information provided by hierarchical naming structures.

To assess the correlation between directory names and file types, project staff created a list of the top 500 directory names drawn from a data set of 1,044 FTP sites. This list was manually reviewed, a subset of "meaningful" directory names was derived, and major categories of file types were established. Project staff obtained sample files from selected FTP sites containing key directory names in the file hierarchy. The files were examined, and correlation between file types and directory names was noted. This process was repeated, and the list of directory names was refined.

The list of directory names served as the basis for a dictionary of regular expressions (rules allowing the matching of various combinations of upper- and lowercase characters, variant spellings, and partial character strings). This process was repeated to refine the dictionary. Using the data dictionary, project staff developed software to parse directory path names, thus enabling automated classification of files.

Summary Analyses

Two random samples of 20 FTP sites were extracted from the total then available from the archie listings file. The data were then parsed, yielding the classification and statistical analyses shown in Figure 1. The percentage of file types for the two samples was very similar, giving a measure of confidence in the algorithm.

The rules for categorizing files were changed based on an analysis of the results of the two 20-site samples, and the categories were adjusted. The final categories were as follows: system code—software, including operating system software, associated with the administration of a computer system; source code—software programs and applications or their components; news—archives of newsgroups and discussion lists arising from group electronic mail transactions; text—files containing or intended to produce, in conjunction with other software applications, a textual document; PC (personal computer)—software applications identifiable as intended for use on personal computing systems; data—raw information, often numerical; images—representations of visual objects, to be

used in conjunction with image-viewing software; games—software applications primarily used for entertainment; executable—compiled files directly executable by a user or an associated software application. The file categorization program was run against all 1,044 sites. The results of this analysis appear in Figure 2.

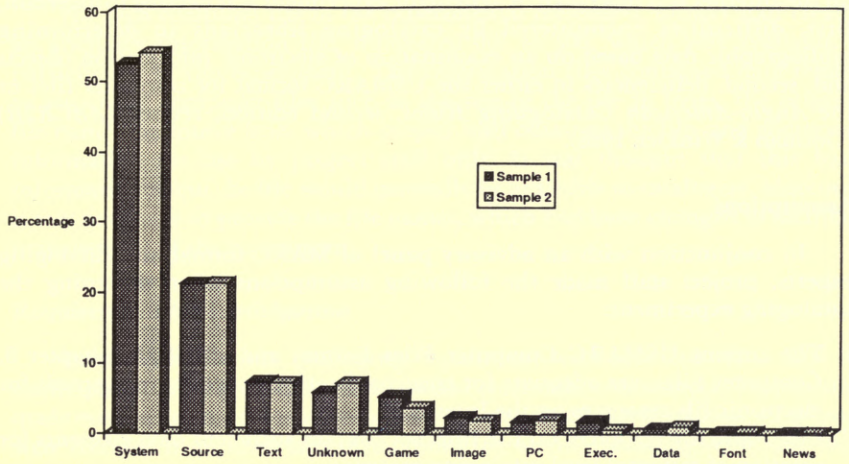

Figure 1. Comparative analyses of two 20-site samples

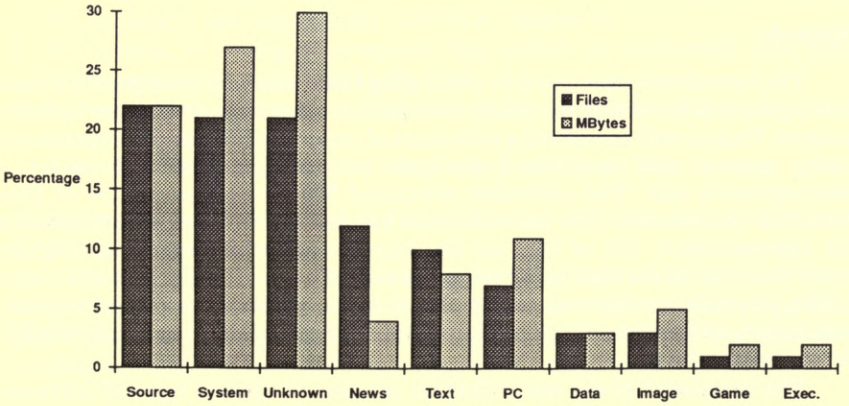

Figure 2. Number and size of files by type (1,044 sites)

CATALOGING EXPERIMENT

Having obtained a statistical overview of FTP sites and their contents, project staff sought to discover the problems, both theoretical and applied,

associated with creating and using machine-readable cataloging records for remotely accessible electronic information objects.

Current MARC (MAchine-Readable Cataloging) records enable the creation, exchange, and subsequent use of machine-readable descriptive cataloging data for a wide spectrum of media, including books, serials, audiovisual materials, maps, musical scores, realia, and computer files. However, the applicability of the MARC cataloging model to the types of electronic information existing on the Internet was unknown. A systematic experiment was devised to ascertain, first, difficulties encountered by cataloging librarians in determining bibliographic data based on an examination of electronic information objects and, second, deficiencies in either the USMARC format for computer files or the *Anglo-American Cataloguing Rules*, second edition, revised (*AACR2R*) (Gorman & Winkler, 1988).

Assumptions

In conjunction with an advisory panel of MARC format and cataloging experts, project staff made the following assumptions when designing the cataloging experiment:

1. The current USMARC Computer Files Format and *AACR2R*, chapter 9, Computer Files, are adequate for creating descriptive cataloging records for electronic file resources on the Internet.
2. Electronic file resources on the Internet contain sufficient data elements for creation of minimal level cataloging records.
3. Catalog records can provide essential access information for electronic file resources on the Internet by incorporating selected fields from the USMARC Format for Holdings and Locations.

Methods

The hypotheses were tested by cataloging librarians who attempted to create catalog records for actual electronic information objects obtained from Internet sources. Project staff assembled a collection of computer files for this experiment using manual and automated methods to minimize bias and ensure a heterogeneous mix. The collection contained 300 files representative of files found on the Internet. The test collection focused on text files, which account for approximately 10% of files on the Internet but which comprise slightly more than half of the files in the test collection. The text files include electronic books, journals, newsletters, poetry, essays, lyrics, guides, lists, papers, reports, legislation, and a range of informal, unpublished materials including USENET newsgroup archives.

The remainder of the test collection consists of various types of software and data files such as source code, programs, games, images, and font files. These files are present in the collection in roughly the same proportions as they exist on the Internet, as determined by our earlier sampling; thus, source files predominated.

Items were not selected for inclusion in the test collection based on their merit as candidates for cataloging. The experiment was intended to test whether the files *could* be cataloged, not whether they *should* be cataloged.

To ensure sufficiently rigorous testing and to minimize difficulties encountered by any single cataloger, each computer file was to be cataloged by three different catalogers.

The 300 experimental files were numbered 001 to 300 and randomly sorted into 10 groups of 30 files each. This process was repeated to yield 30 groups of 30 files; each file occurring in 3 different randomly sorted groups. If each file were cataloged, this would yield 900 catalog records (3 × 300).

For each computer file in the experimental collection, project staff created an ancillary information file, numbered 001 to 300, which was available to participants in the experiment. The information files contained data related to associated computer files such as the size of the file in bytes, the original file name, the source from which project staff obtained the file, and additional information for use by project staff only. It was thought that this basic information about the file would generally be available to catalogers. Moreover, it was necessary to provide the file names, which had been changed to facilitate management of the experiment.

Requirements for Participation

Participants for the cataloging experiment were solicited via a Call for Participants, which was posted to several electronic conferences. Project staff made every effort to ensure widespread opportunity for participation. Responses were received from librarians throughout the world, including Australia, New Zealand, and Hong Kong. The published requirements for participation are given below:

1. MLS degree or equivalent, experience cataloging computer files, and a working knowledge of both the USMARC Format for Computer Files and the applicable cataloging rules (*AACR2R*, chapter 9).
2. Willingness to catalog 30 computer files within the three-week time frame of the experiment from May 11-29, 1992.
3. Internet access, although a limited number of non-Internet sites will be selected, if possible.
4. A word-processing system that can produce ASCII text files.

Participation was not limited to OCLC-member libraries, and online access to OCLC was not required.

Thirty-seven librarians responded to the call. Thirty individuals or teams were selected as primary participants; the remainder were considered auxiliary participants. On average, the 30 primary participants had three years' experience cataloging computer files, although experience ranged from one to 12 years.

Experimental Procedures

Participants were provided with guidelines and instructions. Insofar as possible, all communications and file exchanges related to the experiment took place via the Internet. (The Internet facilitated all administrative aspects of this experiment, including collecting and distributing files for cataloging, distributing ancillary documents, and receiving catalog records created by

project participants.) This communication medium was augmented, when necessary, by phone, fax, and U.S. mail.

Each participant was assigned an identification number from 01 to 30 (auxiliary participants were indicated by the letter "a," e.g., 01a). The identification number corresponded to a similarly numbered set of 30 randomly generated sets of 30 numbers from 001 to 300. Using FTP, each participant was to obtain the appropriate set of numbers from an OCLC computer.

Each number in the set corresponded to a similarly numbered computer file to be cataloged. The computer files were named 001 to 300, with the file extension .obj (for "object"), and each associated information file was named 001 to 300, with the extension .info (for "information").

Participants were instructed to retrieve the assigned object and information files from an OCLC computer using FTP. Project staff provided a record template. The record template contained the valid fixed-field mnemonics and variable-field tags for the Computer Files format, with the addition of field 852 from the USMARC Format for Holdings and Locations. Participants were to complete the record using whatever cataloging aids were available to them and submit the completed record to OCLC, again using FTP.

In addition, participants were requested to complete a log file for each item cataloged and to record in this file the number of the object file and the time required for cataloging. Optionally, participants could record comments, suggestions, or problems related to the object, the cataloging rules, or the MARC format.

Experimental Results and Analysis

Of the 300 electronic files in the test collection, one or more bibliographic records were created for 291 (97%). For these objects, a total of 714 (79.4%) records were created; 650 (72%) log files were created for 291 objects.

The bibliographic records created were analyzed automatically and manually. Automated methods determined the occurrence of a particular field, the length of the field, and the degree of similarity among identical fields when more than one record was created for a single object. Any interpretation of results, however, must include the following overarching factors:

1. Although the participants had experience cataloging computer files, they generally lacked experience cataloging electronic files of the sort included in the experimental collection.
2. Some participants lacked experience cataloging serial materials.
3. The participants were unfamiliar with some of the experiment's guidelines, particularly those relating to location, access, and acquisition information, or the suggested guidelines provided inadequate direction or instruction.
4. In some cases, technical problems confounded the cataloging task.

Despite these limitations, the results of this experiment reflect a substantial amount of empirical data provided by competent and experienced professionals. Summary conclusions are presented in the following sections.

Fixed Fields

With the exception of date fields, fixed-field data consist of single-character codes. Records exhibited a high degree of similarity among these fields with

the exception of "country," "dates," and "encoding level." Despite the likelihood that fixed-field elements will be coded similarly, the number of fixed fields included in the records ranged widely. This indicates that catalogers disagreed as to whether a field should be included in the record.

The "dates" field, while present in most records, exhibited dissimilarity among records for the same object, which may indicate difficulty in determining dates related to computer files.

Variable Fields—Authors and Titles

Variable fields, due in part to the longer text strings they contain, exhibit greater dissimilarity among records for the same object. Two key fields, 100 (Main Entry Heading, Personal Name) and 245 (Title Statement), exhibited the most similarity; however, records do indicate some difficulty determining or recording the authors and titles of computer files. Some records did not contain a 245 field, which is a required field in a bibliographic record. Minimally, every record should have a title field, which may contain the file name itself (*AACR2R*, 9.1B3).

For the 100 field, when two or more catalogers recorded a personal name as a main entry heading, the similarity of the entries was high. However, often one or more records for the same object would not contain a 100 field, although it had been supplied by at least one cataloger. In addition, the overall occurrence of 100 fields was low, appearing in only 18% of all records. This may indicate that this information is lacking or difficult to identify in the information objects.

Notes Fields

Not surprisingly, notes fields were thought to be valuable, occurring in 78% of all records, but the contents of the fields varied greatly. Much of the information provided in the 5XX fields related to subscription or acquisition information, which assumes even greater importance in an electronic, networked environment. The low similarity score may indicate the need for additional fields or subfields to record information now relegated to notes fields.

Location and Access

Two fields could meet the need for expressing location, access, and acquisition information and thereby lessen the reliance upon free-text notes fields: 037 (Subscription Address) and 85X (Electronic Location and Access). With format integration, field 037 will subsume field 265 (Source for Acquisition). This field could record subscription information and instructions, which is particularly important for electronic serial publications. A new 85X field, modeled after the existing 852 field (Location/Call Number) could provide coded location and access information. (For this experiment, field 852 was used for electronic access.)

For acquisition and electronic location and access information, accurate coding of this information is essential. However, coding problems were evident in both the 037 and the 852 fields among records for the same object. In addition, electronic location and access information occurred in only slightly more than half the records. To be effective, all records should contain location and access information for remotely accessed electronic files.

RECOMMENDATIONS

The findings of this project reveal aspects of electronic information objects available via the Internet, provide a taxonomy of file types available via FTP, and, through repeated application under test conditions, provide a substantive body of data on the suitability of conventional methods for providing bibliographic description and access for Internet information objects.

Clearly, the Internet is a rapidly growing environment that facilitates and encourages the creation and dissemination of electronic information objects. As network access broadens, data storage costs drop, and bandwidth increases, the problems of discovering, accessing, and using information on the Internet will likely compound in the absence of additional information management tools and services.

Experimental methods and systems such as WAIS, Gopher, and archie begin to address the problems of network information management; continued research and development of these and other systems are warranted at this early stage of network development and deployment.

To date, remote access, electronic information objects and network information management systems are not well integrated within existing library infrastructures. The reasons for this are many, among them: lack of Internet connection, lack of awareness of electronic information, perceived lack of value of electronic information objects, and the difficulty of locating, accessing, and using electronic information.

Libraries must continue to provide value-added services to the nation's growing body of electronic information objects, systems, and services. While aspects of this electronic information collection—mutability, lack of fixity in a medium, remote accessibility—require adjustments in procedures for cataloging description and access, they do not argue for the abandonment of existing methods. To the contrary, the value of the nation's existing infrastructure of libraries, library systems, and local, regional, and national union catalogs must be leveraged for the information needs of the future.

Libraries stand ready today to begin or to continue the process of providing bibliographic control for remotely accessed information objects. The value of information on the Internet varies widely, and its usefulness is often best determined by the individual user. However, even as not all print materials are collected by all libraries, neither should all electronic files be cataloged. Experience gained in the course of this project indicates that the body of formal, published information is actually rather small when compared with the amount of information available.

As with print and other media, libraries can continue to provide the value-added service of selecting materials for description and access, or inclusion in a collection, whether it be local or remote and dispersed.

As a practical starting point, libraries could create bibliographic records for electronic information objects produced by the faculty or staff of their home institutions. By creating such records, complete with location and access information, libraries help assure broad awareness and access to the work products of the institution. When contributed to a searchable union catalog, these records become widely available.

As a second step, libraries could create records for materials requested and obtained on behalf of patrons, if such records do not already exist. Following the bibliometric wisdom that the best indicator of a subsequent use of an object is an initial use, libraries could contribute to a growing catalog of resources, regardless of whether the library chooses to obtain the file for local holdings.

From this reasoning and the experience gained through this project, the following recommendations are offered:

1. Implement the creation of MARC records for remotely accessible electronic information objects.
2. Monitor the use effectiveness of records created for providing description and access information.
3. Extend cataloging rules and formats to include interactive network systems and services.

To achieve recommendation 3, further examination of the nature of electronic information systems and services is warranted.

The proposals put forward herein do not address all outstanding problems. For example, while electronic description and access are clearly needed, methods of assuring immutable location and access indicators would extend the value of bibliographic records.

LIBRARIES AND THE INTERNET

The Internet facilitates unprecedented connectivity among users and the dissemination of electronic information as never before possible. By every measure, whether you consider the number and types of information objects, the speed of their transmission, or the worldwide breadth of their distribution, computing technologies and high-speed, wide area networks have changed radically the ways in which information is created, stored, and disseminated.

While underlying information technologies have undergone recent and rapid change, the need to preserve and organize information for efficient access, the types of services historically provided by libraries and information professionals, remains an age-old constant.

With the deployment of a new national information infrastructure, such as the National Research and Education Network, a chief concern should be the integration of the Internet and the existing (and enormous) infrastructure of the nation's libraries, library systems, and local and union catalogs. Using existing record formats and cataloging rules, modified as necessary, libraries can begin immediately to provide improved description and access for an important segment of electronic information objects available via the Internet.

Because many libraries participate in national union catalogs, and because the Internet provides widespread access to individual libraries' catalogs, these catalog records created for remotely accessed electronic files would be widely, and immediately, available to information users worldwide. Using existing library systems in this way adds value to the electronic information objects

(through improved description and access information) and adds value to the Internet itself by leveraging for the benefit of all users the familiar information services provided by our nation's libraries.

ACKNOWLEDGMENTS

The OCLC Office of Research gratefully acknowledges the support of the U.S. Department of Education, Office of Educational Research and Improvement, Library Programs, and the contributions of the following, without whose help this project would not have been possible: Glee Cady, Diane Kovacs, Ann Okerson, and Peggy Seiden. Special thanks to our Internet Resources Cataloging Experiment Advisory Committee—Priscilla Caplan, Rebecca Guenther, William W. Jones, Jr., Nancy B. Olson, and Glenn Patton—and to the many volunteer participants in the experimental cataloging portion of this project. Special thanks to Peter Deutsch and Alan Emtage for providing access to the archie listings database.

REFERENCES

Alberti, B.; Anklesaria, F.; Lindner, P.; McCahill, M.; & Torrey, D. (1992). *The Internet Gopher protocol: A distributed document search and retrieval protocol.* University of Minnesota, Microcomputer and Workstation Networks Center (ftp boombox.micro.umn.edu; directory: pub/gopher/gopher_protocol; file: protocol.txt).

Barron, B. (1992). *UNT's accessing on-line bibliographic databases.* University of North Texas (ftp ftp.unt.edu; directory: library; file: libraries.txt (ASCII) or libraries.wp5 (binary for WordPerfect 5.1 file).

Deutsch, P. (1992). Resource discovery in an Internet environment—The archie approach. *Electronic Networking: Research, Applications and Policy, 2*(1), 45-51.

Frey, D., & Adams, R. (1991). *!%@@:: A directory of electronic mail addressing and networks.* Sebastopol, CA: O'Reilly.

Gorman, M., & Winkler, P. W. (1988). *Anglo-American cataloguing rules* (rev. 2nd ed.). Chicago, IL: American Library Association.

Kahle, B., & Medlar, A. (1991). An information system for corporate users: Wide area information servers. *Online, 15*(5), 56-60.

Kehoe, B. P. (1993). *Zen and the art of the Internet: A beginner's guide* (2nd ed.). Englewood Cliffs, NJ: PTR Prentice Hall.

Kovacs, D. K. (1991). *Directory of scholarly electronic conferences.* (ftp ksuvxa.kent.edu; directory: library; files: acadlist.file1, acadlist.file2, acadlist.file3, acadlist.file4, acadlist.file5, acadlist.file6, acadlist.file7).

Krol, E. (1989). *Hitchhikers guide to the Internet.* Network Working Group Request for Comments 1118. (ftp nis.nsf.net; directory: /documents/rfc; file: rfc1118.txt).

Krol, E. (1992). *The whole Internet: User's guide & catalog.* Sebastopol, CA: O'Reilly.

LaQuey, T. L. (Ed.). (1990). *The user's directory of computer networks.* Bedford, MA: Digital Press.

Lincoln, B. (1992a). *Wide Area Information Servers (WAIS) bibliography.* Menlo Park, CA: Thinking Machines Corp. (ftp quake.think.com; directory: pub/wais; file: bibliography.txt).

Lincoln, B. (1992b). Wide Area Information Servers (WAIS) bibliography. *Information Standards Quarterly, 4*(3), 13-15.

Lottor, M. (1992). *Internet growth (1981-1991).* Network Working Group Request for Comments 1296 (ftp nis.nsf.net; directory: /documents/rfc; file: rfc1296.txt).

Malkin, G., & Marine, A. (1992). *FYI on questions and answers: Answers to commonly asked "New Internet User" Questions.* Networking Working Group Request for Comments 1325 (ftp nis.nsf.net; directory: /documents/rfc; file: rfc1325.txt).

Marine, A. (Ed.). (1992). *Internet: Getting started.* Menlo Park, CA: SRI International.

National Science Foundation. Network Service Center. (1989). *Internet resource guide.* Cambridge, MA: BBN Systems and Technologies Corporation (ftp nnsc.nsf.net; directory resource-guide).

Nickerson, G. (1992). Getting to know Wide Area Information Servers. *Computers in Libraries, 12*(9), 53-55.

Quarterman, J. S. (1990). *The matrix: Computer networks and conferencing systems worldwide.* Bedford, MA: Digital Press.

Scott, P. (1992). HYTELNET as software for accessing the Internet: A personal perspective on the development of HYTELNET. *Electronic Networking: Research, Applications and Policy, 2*(1), 38-44.

St. George, A., & Larsen, R. (1991). *Internet-accessible library catalogs & databases.* Albuquerque, NM: University of New Mexico. (e-mail list-serv@unmvm.bitnet; message: GET LIBRARY PACKAGE.)

Strangelove, M.; Okerson, A.; & Kovacs, D. (1992). *Directory of electronic journals, newsletters and academic discussion lists* (2nd ed.). Washington, DC: Association of Research Libraries.